



The Official  
Statistics System

# Uniques and Disclosure Control Design <sup>1</sup>

L. Fraser Jackson.  
*Emeritus Professor, Victoria University of Wellington*

*For correspondence: [fraser.jackson@vuw.ac.nz](mailto:fraser.jackson@vuw.ac.nz)*

*A paper prepared for Statistics New Zealand under an OSDRAC contract*

---

<sup>1</sup> I appreciate the support of Statistics New Zealand for this project. The views expressed are those of the author and do not represent the views of Statistics New Zealand. The author owes much to the assistance of Lisa Corscadden who provided helpful comments and made time for discussion whenever asked, and Irene Zeng who provided insightful comment and assisted with the preparatory data analysis. There are many points where their comments have clarified the text. This paper provides a more comprehensive treatment of issues discussed in Jackson, Corscadden and Zeng (2005) and some applications of them.

## *Executive Summary*

The statistical disclosure control (SDC) literature has traditionally focused on tools to prevent disclosure once a potential disclosure problem has occurred with little attention to describing the circumstances which generate sensitive cells and ways of avoiding them.

The objective of this paper is to study the occurrence of sensitive cells in tables of count data and provide some guidance on the SDC rules which are necessary to help protect the anonymity of respondents.

The main risk occurs when cells contain entries of a one or a two. These cell entries cannot be published without some disclosure control. Cells with ones are referred to as uniques.

The Multinomial Distribution  $MN(n,p)$ , with  $n$  the sample size and  $p$  the vector of category probabilities, is a simple widely used model for the numbers in table cells. For sample sizes and probability vectors which occur commonly in statistical practice it can generate large numbers of cells containing uniques.

We study in detail tables of data generated at meshblock level for variables which are either published or could be requested by Census users. For these tables we show that the Multinomial distribution provides a satisfactory model for studying the behaviour of uniques.

A first approximation using just the national category probabilities provides a good initial framework, but we show that adding a means of allowing for spatial variation in the probabilities improves the fit so that very good estimates of the total number of uniques observed at different meshblock sizes are obtained.

Our empirical analysis shows that there are large numbers of uniques for commonly occurring situations. Hence it is important to find systematic ways of providing measures to inform a statistical agency about disclosure control risks.

In view of difficulty of knowing the information available to possible intruders, the only practical way of assessing disclosure control risks is to measure the incidence of events which would create a disclosure control problem if some SDC tools were not applied.

We suggest two measures. The first is the probability that a table cell entry is a unique. This measure helps control the probability that a person in a minority group in the population is identifiable. The second is the probability that in a table the entry for a person from the population appears as a unique in that table. This measure helps assess the overall risk and low values make intrusion a less rewarding activity to an intruder since fewer persons are potentially identifiable.

### The Cell Probability measure

In the tables studied for the 2001 Census there were 8.38 million cells. Of these about 350,000 cells were situations where there was a probability of 50 percent or more that the cell would contain a unique. Without adequate SDC, information from these individuals would be revealed. Almost half of all cells had a probability of 10 percent or more of containing a unique. This places a huge burden on the SDC procedures.

It is suggested that suitable targets for the cell probability measure may be in the range from 10 to 2 percent.

These would impose significant restrictions on the values of  $n$  and  $p$  which can be permitted. It would imply global recoding to a more aggregated classification for many of the smallest categories, but the extent of the recoding required would depend on selection of new publication units with a larger minimum size.

The empirical study shows that significant reductions in the maximal risk for particular cells can be achieved by forming new publication units which have a clear size lower bound.

## The Person Probability measure

From the perspective of the second measure, there were 734,341 cells which were uniques in a set of tables with a total person count of 104.6 million. This still gives the potential for it being a significant proportion of the population, of 3.74 million but on average across the tables gives a probability of 0.7 percent of person entries being uniques.

To reduce the potential value to an intruder of creating additional links with the data it is important to reduce this measure. It will obviously vary greatly across tables. We have only calculated it at an aggregate level for the set of 28 tables studied.

In that framework the target might be to reduce the total number of unique cells to a sufficiently low proportion of the number of persons in the population. Targets at the low percentages of the population will be difficult to achieve.

The Person Probability measure also provides a framework for discussion of the impact on the risks of adding or deleting a table from a set of published tables.

## Suggested actions

It would be highly desirable from the perspective of both risk measures to move to the system in both Australia and the UK of having a very much larger minimum size for data publication units. A minimum of at least 100 might be at half a target size of 200 or more.

Careful attention should be given to classifications to minimize the number of categories and especially to avoid categories with small proportions of the population wherever that can be achieved without significant loss of information.

It is very difficult to take the analysis of risk further because little is known about the performance of alternative SDC measures on source tables which contain uniques when the publication tables are subject to alternative methods of attack. This clearly merits further research.

Maximising the category probabilities is obviously important in reducing risks. While much is known about problems such as Simpson's paradox, the best ways of forming aggregates in multidimensional data sets does not appear to have been studied extensively. It is central to developing publication methods which minimise information loss at the same time as minimising disclosure risks.

5 January, 2005

## **Introduction**

The statistical disclosure control (SDC) literature has traditionally focused on tools to prevent disclosure once a potential disclosure problem has occurred with little attention to describing the circumstances which generate sensitive cells and ways of avoiding them. The objective of this paper is to study the factors which lead to numbers of sensitive cells and provide some guidance on rules which help reduce those numbers and the reliance which must be placed on SDC to preserve the anonymity of data providers.

In the paper we will focus on cells which contain an entry of one. We refer to such entries as uniques, since a single person has the characteristics for that table entry. Uniques are not the only source of sensitive cells, and cells with entries of two and three are also commonly included as sensitive since they may permit a person to identify another person with the same characteristics. The occurrence of cells with an entry of a two or a three is closely related to occurrence of ones, and while there are some additional considerations many of the issues remain the same. We therefore concentrate on uniques.

This paper examines tables of counts of persons from a census. It arose because in a previous study of some sets of univariate tables for small area statistics there were up to 25 per cent of the cells which were uniques. The problem of sensitive cells was spread very widely through these tables, and the ability of statistical disclosure control tools to handle the problem had to be questioned. In Jackson, Corscadden and Zeng(2005) we report briefly on further examination of the issues raised in the previous work by Zeng. The problem is exacerbated when parts of the tables are particularly sparse. With continuing improvements in statistical techniques and in computational tools, more avenues of attack become available so a data agency needs to take increased care to protect respondents. Steps to control the number of sensitive cells prior to application of SDC therefore become very attractive. The use of global recoding is often cited as an example of a method which may be necessary (see for example Willemborg and de Waal(1996)(2000).) but there is no quantitative assessment of its effects.

Studies of uniques by Skinner and Holmes(1992), Elliot, Skinner and Dale(1998) and Elliot(2001) have examined models for the incidence of uniques in tables and microdata. Other earlier studies of the frequency of frequencies in tables are in Bishop et al. (1975). The Skinner and Holmes approach describes and estimates a compound poisson-lognormal distribution to estimate the number of uniques. Unfortunately their method requires the table to estimate the model parameters. We consider it more useful to explore the range of empirical situations which arise, and endeavour to find those features which are necessary to model the number of uniques which will occur in a table.

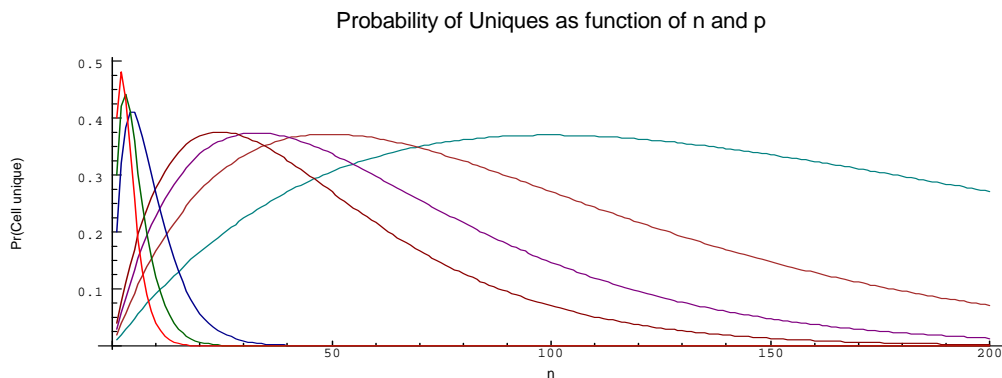
This paper provides more detail and examines policy issues which the results raise. We outline a simple model, and then explore the behaviour using its parameters. Elaboration of the ideas shows the necessity of introducing geographic variation in parameters of the model to further improve the fit and doing so confirms the value of the Poisson approximation we use. The paper concludes with an examination of the effects of some alternative publication design rules and a range of suggestions which would help reduce the incidence of situations to which SDC must be applied.

These ideas would be even more useful if the risk of breaking an SDC tool had been studied for a wide range of alternative tools. In Jackson (2004) we explored the risks associated with random rounding over a part of its range of application. Many aspects of the performance of SDC tools are not well understood creating an important uncertainty for statistical agencies.

## **The multinomial model**

The multinomial model provides a framework for considering the problem. For each of the  $n$  individuals in a geographic unit, there is a vector  $p$  of probabilities associated with each of the categories in a classification. The number of observations in each category can be modelled as a vector  $X \sim MN(n,p)$ . This assumes independence between observations of individuals which is

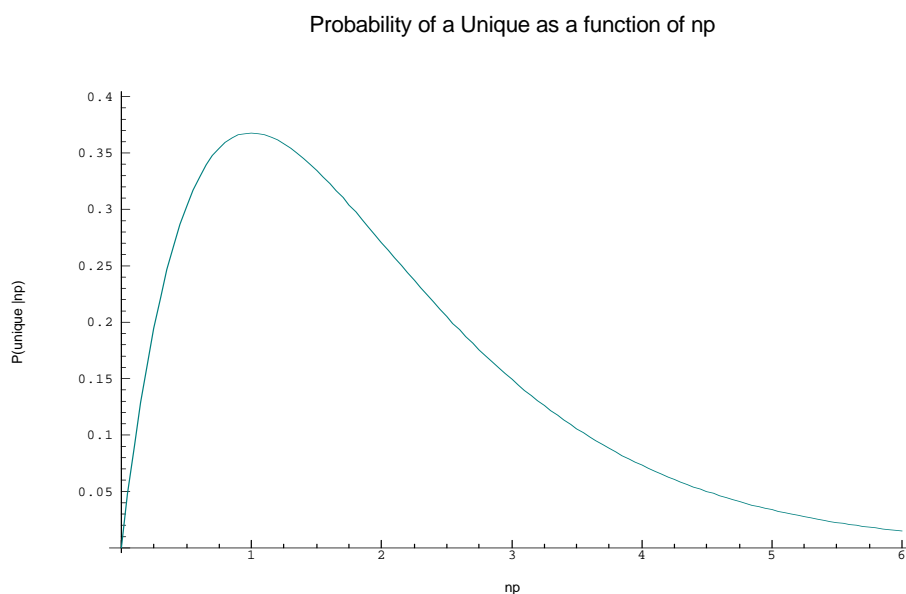
clearly not met, but we will maintain it as an adequate first approximation to the behaviour. It is well known that the marginal distribution for each category can be analysed as a binomial with the expected number of uniques in the category  $E[\mu_1] = np(1-p)^{n-1}$  where  $p$  is now the category probability. Over the range of values of interest this is adequately approximated by the Poisson form  $E[\mu_1] = npe^{-np}$ . These expressions can be summed across categories to obtain the expected number of cells with uniques in a table. Note that the expression can be parametrised with  $a = np$ . Changes in  $n$  have exactly analogous effects to the corresponding proportional changes in  $p$ . The maximum expected number of uniques occurs with  $a=1$  and the proportion of persons who are uniques is a continuously decreasing function of the unit size  $n$ . For a multinomial, this model generates a wide range of behaviours depending on the actual distribution of the values in the vector  $p$ . Figure 1 illustrates the case  $p = (0.4, 0.3, 0.2, 0.04, 0.03, 0.02, 0.01)$



**Figure 1. The expected proportion of uniques in a category as a function of n**

The curves have peaks at larger  $n$  as the probability decreases. Note how flat the curves are for small probabilities and how rapidly they fall towards zero with increasing  $n$  when  $p$  is large. The total expected number of unique cells need not be monotonic in  $n$ , and the counter intuitive behaviour of increasing numbers of uniques as  $n$  increases over a wide range of values is possible.

While we have noted the probability of a unique can be parametrised by  $a = np$  there are important features of this simple form. The sample size  $n$  is integral and must be greater than 1. The values of  $p$  are restricted to the interval  $[0, 1]$ .



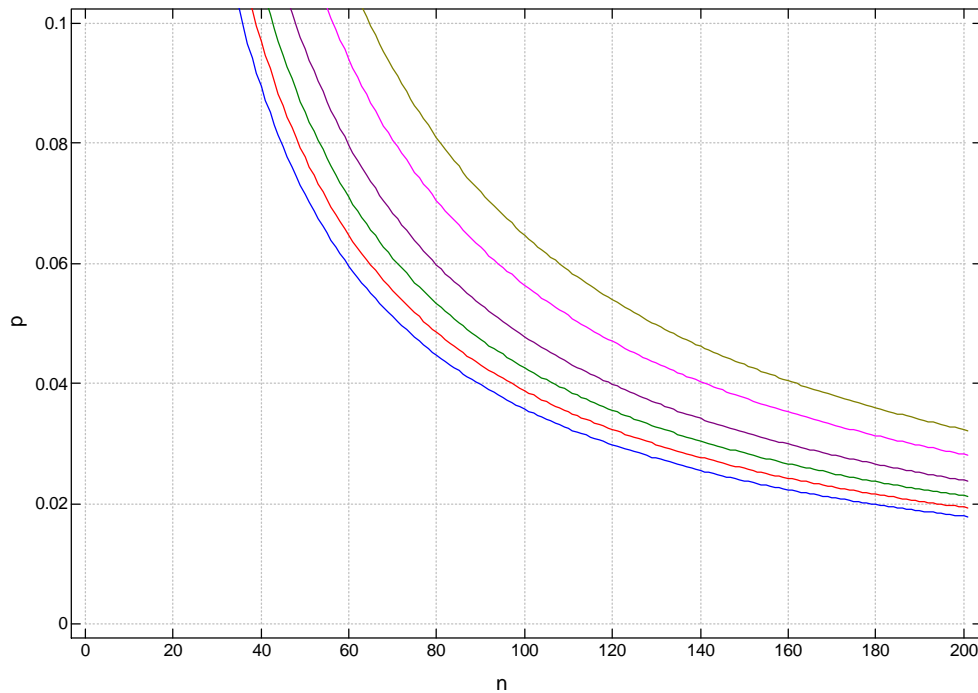
**Figure 2. The expected number of uniques in a category as a function of np**

Figure 2 gives the changing probability of a unique as a function of  $a = np$ . For any given  $p$ , changing  $n$  moves the expected probability of a unique along the curve. It falls to 0.1 when  $np=3.6$ , 0.05 when  $np = 4.5$  and 0.01 when  $np$  reaches 6.5. These figures show that if a statistical agency wants to attain less than one cell in 20 which contain a unique they need to have  $np$  greater than 4.5. Such a level imposes significant constraints on the possible combinations of  $n$  and  $p$ . If the minimum publication unit size is 100 then the minimum category proportion needs to be 0.045, a value which is above the median national category proportion for the set of variables in our empirical studies.

Cell probabilities are not the only way of assessing risk. A cell probability of a unique gives the probability a geographic unit contains a unique. SDC aim to make it very difficult to actually determine if there is a unique or not, but it is clearly undesirable to have a high proportion of cells with uniques as it opens more means of attack. If there is a unique then there is the further question of whether they are identifiable. At the meshblock level a unique or even a two in a table may enable an intruder with sufficient other information to make a re-identification with a high probability of being correct. As noted earlier the expected number of cells with a two is important. It is greater than the number of uniques whenever  $np > 2-p$  so applies over a wide range of situations. This implies there are at least three times the number of persons in sensitive cells for each of the uniques given the  $np$  combinations in the previous paragraph.

In addition to the cell probability we need to consider the probability that a particular person is a unique. Within a category the ratio of the expected number of uniques to the expected number of persons is  $(1-p)^{n-1}$  and this reaches 1 percent when  $np = 4.5$  and  $n = 100$ .

Fortunately most persons are in categories with large  $p$  values. For such categories Figure 1 shows that the expected number of uniques falls rapidly with increasing  $n$ . If there are few categories and their minimum size approaches probability of 0.1 then an area unit size of 50 would exceed  $np = 4.5$  for all categories. However as soon as there are category probabilities below this level the expected number of uniques quickly becomes a very flat curve as  $p$  falls. For categories with small  $p$  controlling minimum area unit size is not adequate for controlling the proportion of uniques. You need a unit size of 1000 when  $p = 0.0045$  and we will see that many categories have smaller  $p$  in our study.



**Figure 3. Contours of equal probability of a unique as a function of  $(n,p)$**

We note the ubiquity of uniques. They can occur for all  $p$  and all  $n$ , and it is useful to plot the level curves for  $P(\text{unique} | np)$ . Figure 3 shows a range of points for probabilities of 0.1 and less. The probabilities have a maximum along hyperbola  $np = 1$ .

### **An empirical study**

So far we have considered a probability model of uniques. It enables us to make deductions given known values of the parameters  $n$  and  $p$ . If we could observe a large number of samples from a single population it would give predictions we could compare with empirical values. We now turn to assessing how well a model of this type fits the observations for meshblocks.

In New Zealand the meshblock is the primary administrative data collection unit. They are determined having regard to historical usage, physical proximity or accessibility, and workloads of census staff. To study the behaviour of uniques at the Meshblock level we used two data sets. The first was a set of 18 variables with a total of 177 categories from the 1996 census. The second is a set of 28 variables with a total of 232 categories from the 2001 census. The categories are not all distinct, since common factors such as age may affect more than one variable. Such common categories have been left in the analysis as they would appear multiple times in a published table set. Some variables were recoded to remove a category such as ‘Undecodable response or otherwise unknown’ which posed no disclosure risk but this was not always done. One variable was recoded to bring it to the highest level of the classification. Our concern was to have a range of census variables and category proportions within them. The variables used from the 2001 census together with category proportions after any recoding are listed in Appendix 1.

The administrative collection units called meshblocks have been the minimum publication unit in the past, but there is no statistical reason why the minimum collection units and the minimum publication units should be the same. The meshblocks for the 2001 Census vary widely in size with some as small as a single person, the median is 93 and the mean size 103. One had in

excess of 1000 persons. From a statistical perspective it would make much more sense to have units which were similar in size since then observed measures calculated for those units would have more similar statistical properties. The primary grounds for the existing definition is administrative convenience during the collection, and for publication it would make sense to aggregate them to larger units taking account of similarity in economic or social functions. Ten percent of existing meshblocks are 19 persons or less and statistics derived from random rounded information on such samples conveys very little information. The existing size distribution was useful in this study because the large number of small units enabled us to observe summary statistics of interest for a large number of cases.

There will always be a need for minimum collection units and that information should continue to be recorded with the main census data set but we will argue below that they create significant problems as the minimum publication units

### **a. A Homogenous model for $p$**

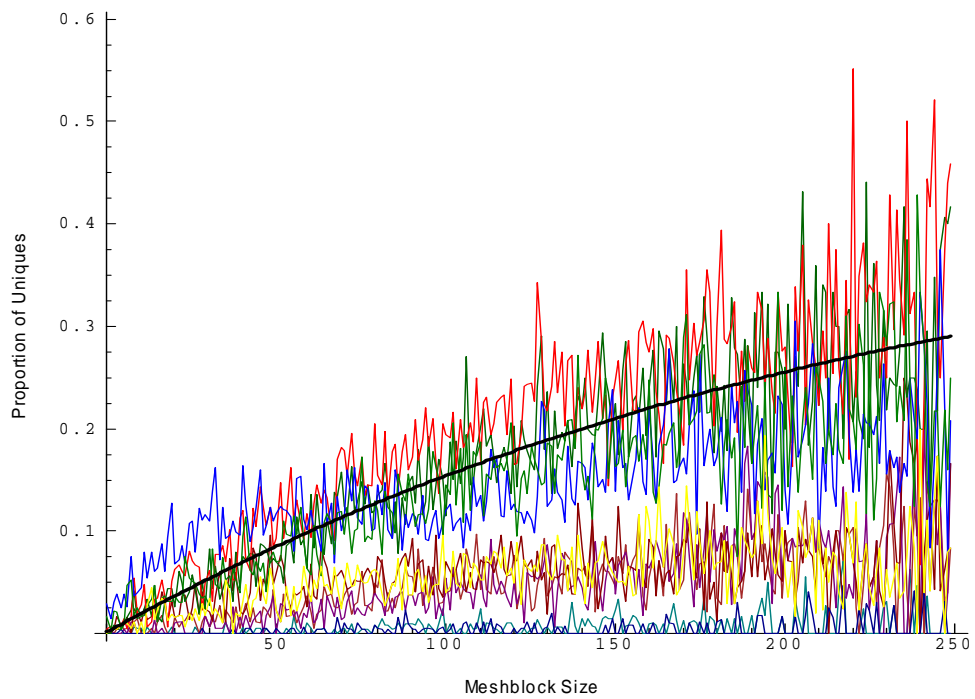
Consider a single category of a variable. We can observe the count of cases for that category for all meshblocks. If they were all drawings from a homogenous population then the MN( $n,p$ ) model described above should provide a good fit to the data.

The value of the model as a first approximation is reinforced by graphs in Figures 4-6. The categories are ranked in ascending order of the value of  $p$  which is the national proportion of the population in the category. The broad similarity of the curves in each graph is striking. The graphs illustrate the observed proportion of uniques for categories in the 1996 census and since duplicates convey no additional information, they were omitted leaving 162 categories. Note that the source variables are ignored in the grouping of the categories. Each trace on the graphs is the proportion of uniques across all meshblock sizes for a single category. These graphs plot the source data to give a visual impression of the density of the distribution of cases and the uncertainty associated with inferences about the relationship which is plotted. They show that the behaviour broadly follows the Poisson model based on the mean proportion for the categories included in each graph, which is the heavy black line on the graph.

Figure 4 for the smallest categories shows a pattern which should be expected, with successive lines of increasing slope as  $p$  increases from the lowest values. At the very lowest proportions the maximum sample size is not sufficient to reach the maximum expected number of uniques. The second group of cases in Figure 5 is of categories about the lower quartile with ranks 41-50. These reach a maximum in the neighbourhood of 50 ( $n=50, p = 0.02$ ) and show a further characteristic pattern with curves dipping below the expected values near the maximum and above it for a range of higher values of  $n$ . The third group in Figure 6 is for ranks 121-130 spanning the upper quartile of the distribution of category probabilities. For these categories the maximum expected number of uniques is reached for low  $n$  and by  $n = 100$  has fallen to much lower levels with a peak observed proportion of uniques about 5 percent. The pattern of deviations about the curve observed in Figure 5 is repeated here.

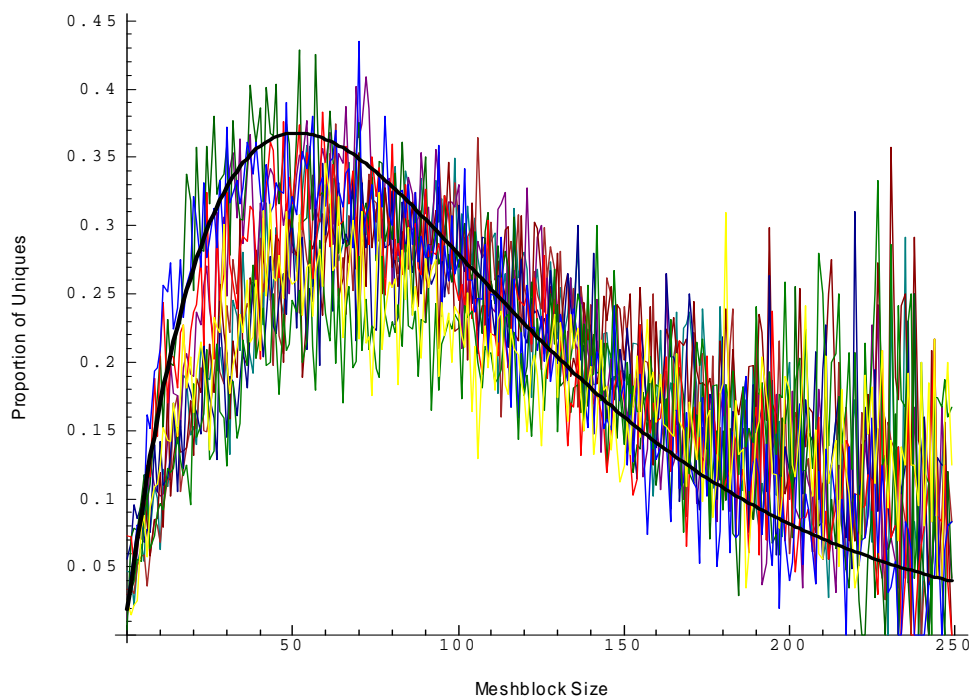
The differences between the numbers in the homogenous model and the observed proportions are sufficiently consistent across many categories that we must consider its source. The whole idea in constructing a meshblock table is to generate some understanding of local differences from a national pattern. Simulation of some cases using a MN( $n,p$ ) distribution to generate the expected proportion of uniques in a category, but with a distribution of values of  $p$  for the geographic units in category gave the characteristic pattern of differences observed in the graphs. In the next sections we consider two alternative simple models which assume meshblocks reflect a wider pattern of variation in the structure of the population with respect to the categories being studied.

Observed Proportion of Unique Cells, Ranks 1-10



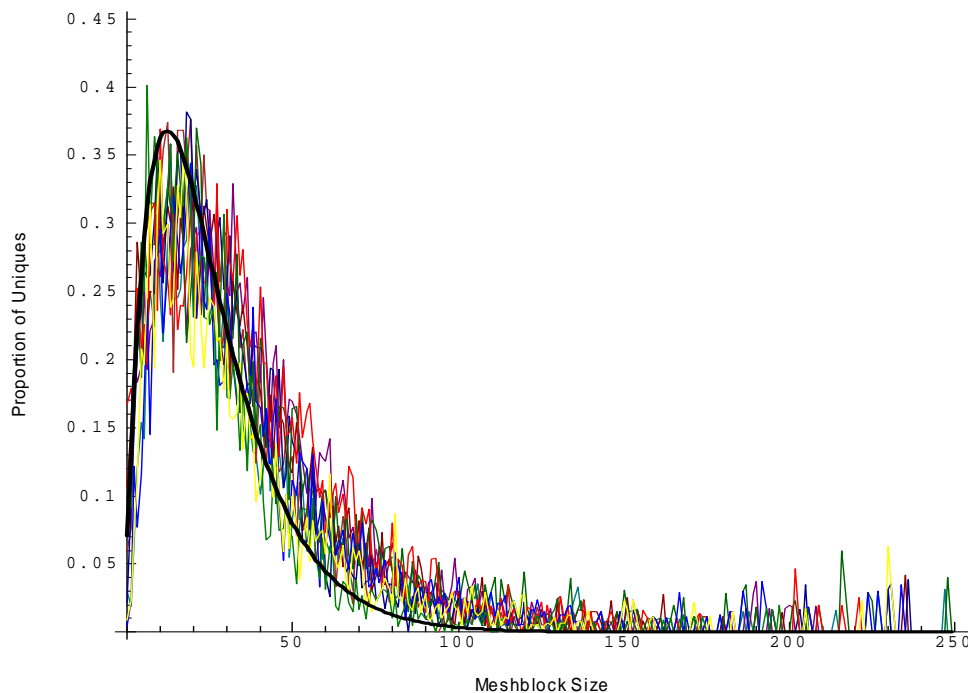
**Figure 4. Trace of Proportion of Uniques for Categories in Ranks 1 - 20**

Observed Proportion of Unique Cells, Ranks 41-50



**Figure 5. Trace of Proportion of Uniques for Categories in Ranks 41-50**

Observed Proportion of Unique Cells, Ranks 121-130



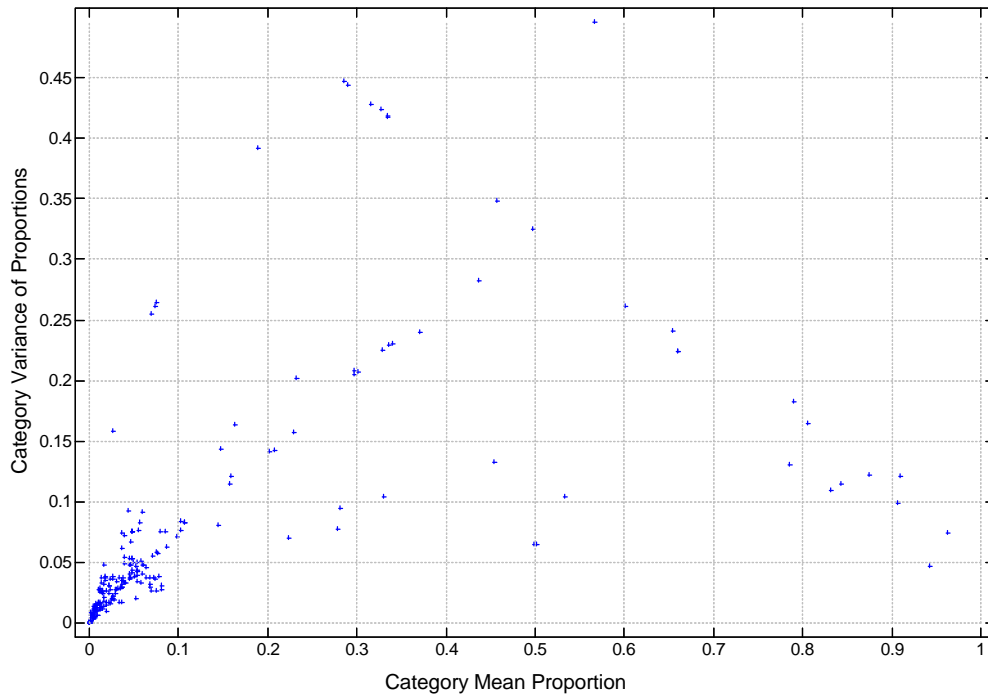
**Figure 6. Trace of Proportion of Uniques for Categories in Ranks 121-130**

### **Introducing Heterogeneity in $p$**

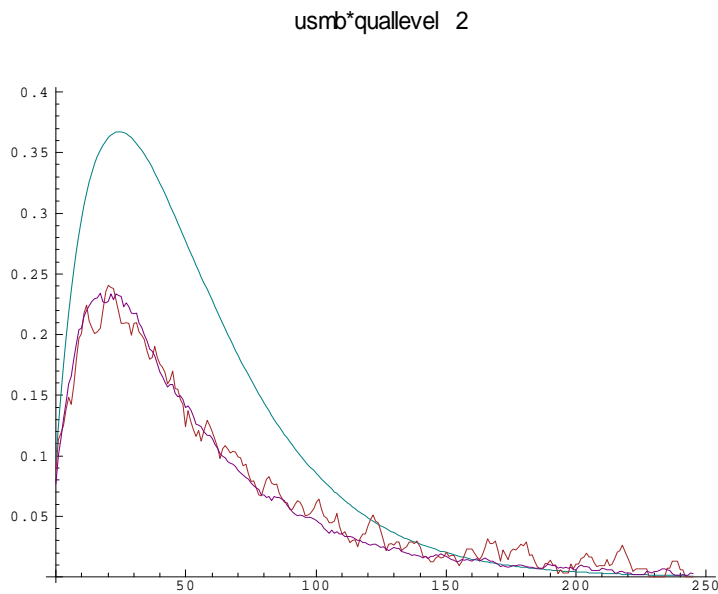
Meshblocks are aggregated to area units. Area units also have a wide range of sizes but we can use them to obtain a more accurate estimate of the proportion of persons in a category than we can obtain from meshblocks. They are also sufficiently large units to get useful estimates of mean and variance of each category proportion. Figure 7 plots the mean and variance for each of the 2001 census data categories. The line shows the expected variance for binomial sampling. There is a clustering of points around this line, but some points depart from it considerably and indicate significant geographical variance as in Lexian sampling, or smaller variances perhaps from Poissonian sampling for some categories. (See Weatherburn (1946)).

A very simple approach to introducing heterogeneity is to construct a model where for each meshblock in an area unit we use the area unit probability for the category in the formulas above. Figures 8-10 give examples of the impact of heterogeneity in the model. With differences in the category probability in different area units, we can calculate the expected number of uniques at each meshblock size. The graphs show that it is a better approximation of the observed pattern of uniques than the homogenous model for the illustrated categories. A similar improvement can be illustrated for many other categories.

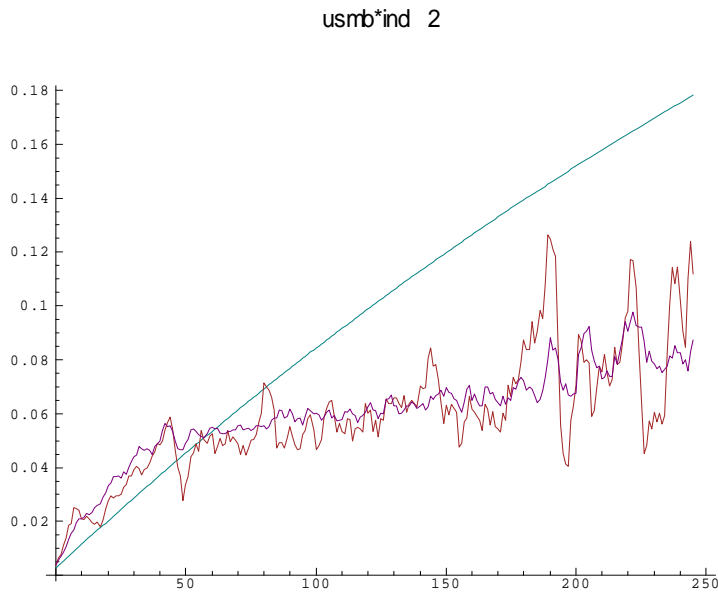
Another approach is to regard meshblocks as the smallest units for observing a process which is changing smoothly over space. While meshblocks are not a single continuous ordering of adjacent areas, in the numbering sequence meshblocks are nearly always adjacent so a simple kernel estimator could be used. We have used a 5 term moving average to replace the value for each meshblock and find that apart from the smallest meshblock sizes it approximates the observed frequency of uniques. Figures 11 and 12 provide some examples.



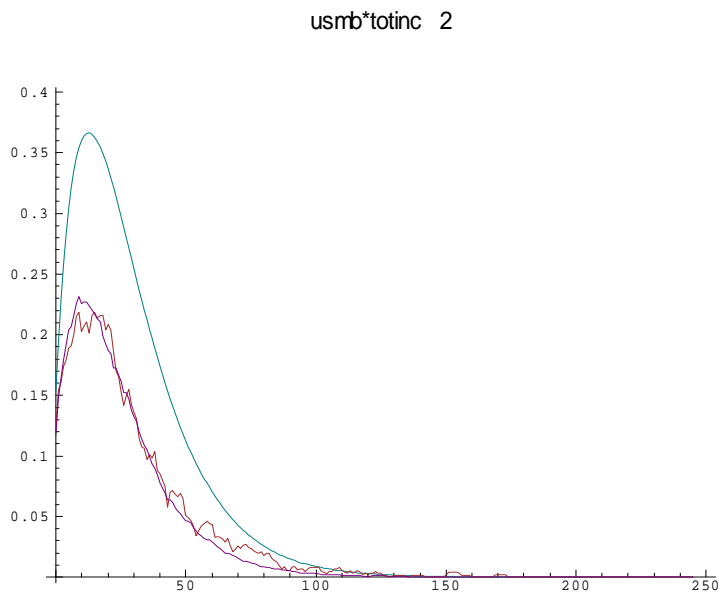
**Figure 7. Category Variance and Category Mean Proportion for 2001 Census categories**



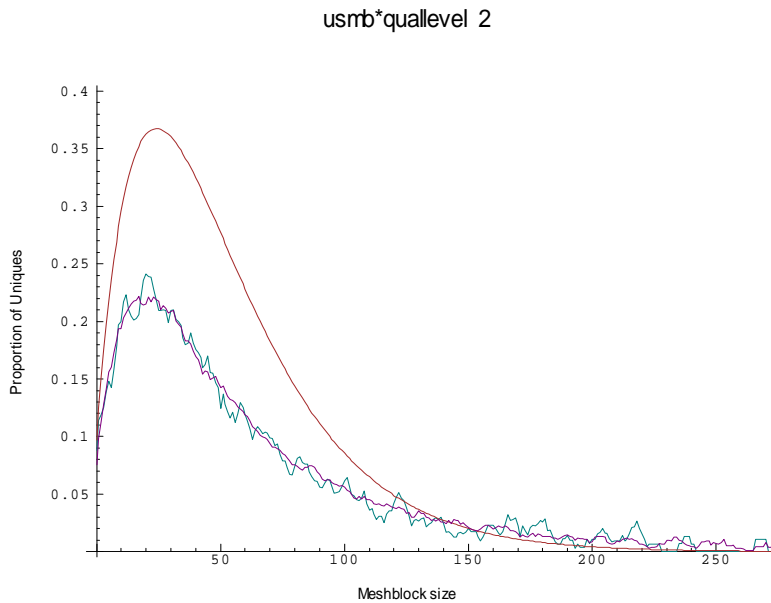
**Figure 8. Observed and Fitted Model Using Area Unit Means (1)**



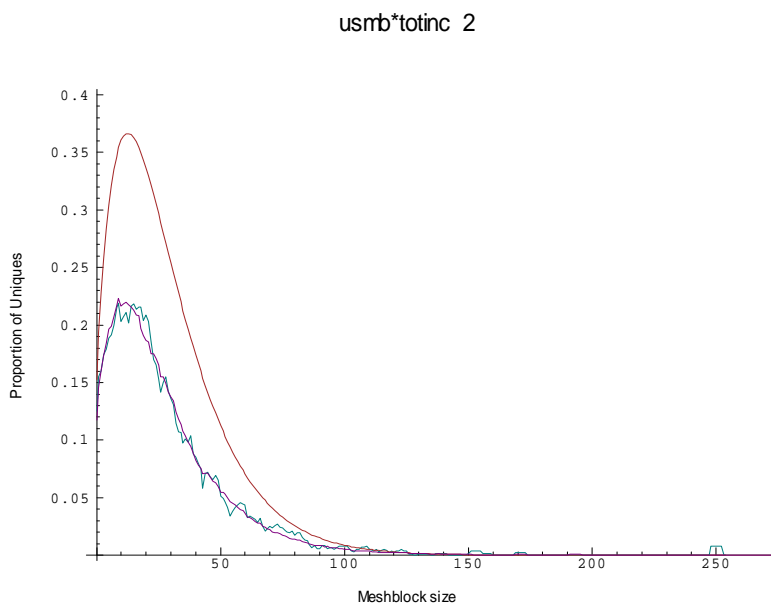
**Figure 9. Observed and Fitted Model using Area Unit means (2)**



**Figure 10. Observed and Fitted Model using Area Unit means (3)**



**Figure 11. Observed and Fitted Model using Local similarity(1)**



**Figure 12. Observed and Fitted Model using Local Similarity(3)**

Figures 8-10 are compelling evidence in favour of the multinomial model as a model for this data when an appropriate adjustment is made for local variation in the values of the category proportions. This however creates a dilemma because the geographic variability of a category may not be known until the data is collected. Figures 11 and 12 where the proportions used are exactly the same as they would be for aggregates of five meshblocks, giving a mean size of about 500 persons, indicates much of local variability is captured at that level of aggregation.

We conclude from this empirical examination that using the Poisson approximation and a mixture distribution for the probabilities of a category can yield a very close estimate of the actual distribution of uniques by meshblock size.

Skinner and Holmes (1992) noted that for some of their distributions it was necessary to allow for the possibility that in the mixture distribution there is a finite weight on the category probability zero. While the homogenous model over estimates the probability of a one, it generally under estimates the probability of a zero. We have not investigated whether this could also arise from

heterogeneity in the category p values across geographic units but it would not be surprising if a mixture with a finite weight on zero provided a better fit for some categories.

While Skinner and Elliot define ‘special uniques’ in a multivariate context their property of having a value of p much less than appropriate values elsewhere in the data space also occurs here with geographic heterogeneity. They are cases where there are very low probabilities compared with the probabilities at places where there are concentrations of people in the category. Applying the general MN(n,p) model then provides evidence that persons at these locations are very likely to be in the tables as uniques.

### **An Aggregate Picture**

To provide a summary description of the whole 2001 data set we formed a table where each row represents a Meshblock size, and each column a variable category. All cells were classified by decile for the number of persons and by decile for the national proportion of persons in the category. Table 1 gives the numbers of uniques for each case. The vertical dimension is in increasing order of size, and the horizontal dimension in increasing order of category proportion. The margins give the maximum size within the decile. Each size decile contains about 3600 meshblocks and each category proportion about 23 categories

**Table 1. Number of Uniques by Deciles of Meshblock size and Category Proportion**

		1	2	3	4	5	6	7	8	9	10
		0.002	0.009	0.015	0.026	0.040	0.052	0.066	0.136	0.358	0.950
1	19	344	2541	4198	7393	10689	11775	13725	14392	10926	6965
2	40	967	6917	9954	14886	17287	16333	14187	11703	4908	275
3	60	1560	9902	12753	16628	15757	14088	9223	5920	3636	85
4	77	1960	12012	14601	16910	13333	11781	6117	3286	3030	56
5	93	2497	13800	15510	16031	11013	9423	4297	1944	2624	25
6	111	2798	14979	15555	14572	9299	7455	3109	1241	2378	30
7	130	3287	15955	15422	12971	7670	5967	2379	767	2206	15
8	155	3668	16614	15109	11219	6333	4432	1811	530	1938	5
9	198	4199	17019	14202	9233	4961	3128	1323	316	1788	1
10	918	5616	16721	12174	6322	3355	1479	908	257	1454	4

Some disclosure control tool needs to be used for each of the uniques in Table 1. Each cell in the table represents approximately 83,000 cells in the source tables so differences in proportions of uniques of less than half a percent (about 400 for Table 1) are highly significant. The peak proportion in the MN model at the individual category level illustrated in Figures 4-8 where many groups of cells have between 30 and 40 percent of uniques is not observed because of aggregation over a wider range of sizes, category probabilities and geographic dispersion. However a large part of the table has approximately 20 percent of the cells which are uniques. Virtually all SDC methods would have difficulty dealing with such large proportions of sensitive cells over a significant region in a table.

**Table 2. Numbers of persons in decile groups for Meshblock size and category proportion.**

decile		1	2	3	4	5	6	7	8	9	10
max		0.002	0.009	0.015	0.026	0.040	0.052	0.066	0.136	0.358	0.950
1	19	846	3533	7999	14862	27876	36231	40363	78255	212947	528276
2	40	2285	12176	26205	48060	89050	108178	128698	252592	666478	1658750
3	60	3291	20210	44934	80557	141860	176241	219273	405192	1118738	2777428
4	77	4442	29110	63514	112712	188378	241986	309210	529256	1520635	3791345
5	93	5500	38619	81680	145493	234472	313483	402073	650795	1882123	4726962
6	111	6588	46620	100471	171614	273889	373952	484393	756910	2270777	5694438
7	130	7280	58195	121969	210205	330843	461077	592744	903461	2659687	6702631
8	155	8667	70442	145981	249304	390509	556049	703795	1075187	3160340	7949182
9	198	10144	88481	180762	311524	482592	700887	886267	1319628	3859563	9710916
10	918	14360	136604	285320	489780	755199	1137320	1402130	2083108	5737699	14311128

There is another way of looking at this data. The cells contain very different numbers of persons. Categories with small proportions have few people in them. A further way of looking at these cases is to consider the number and proportion of persons for each of these decile groups. Table 2 gives the number of persons in each cell of the deciles table. Since there are 28 variables these total to 28 times the national population.

The ratio of the entries in Table 1 and Table 2 shows extreme differences in the risk associated with different categories. Persons in small categories and in small meshblocks have very high probability of being a unique. Whereas there were only 4 uniques in the 14 million in the highest decile on both categories. For these people any disclosure control tool is likely to provide sufficient protection.

Table 3 lists the proportion of uniques expressed as persons per ten thousand for each of the decile groups. Note that for all rows, the number of persons in each cell increases as you move to higher proportions, so most people are in the proportion decile 10. The table shows very clearly that the risk is very much higher for persons in categories with small proportions, and these are often the people of most interest to intruders.

This table again shows that the incidence of uniques needs to be considered as a function of the categories.

**Table 3. Uniques per 10,000 persons by Deciles of Meshblock size and Category Proportion**

		1	2	3	4	5	6	7	8	9	10
		0.002	0.009	0.015	0.026	0.040	0.052	0.066	0.136	0.358	0.950
1	19	4066	7192	5248	4974	3834	3250	3400	1839	513	132
2	40	4232	5681	3799	3097	1941	1510	1102	463	74	2
3	60	4740	4900	2838	2064	1111	799	421	146	33	0
4	77	4412	4126	2299	1500	708	487	198	62	20	0
5	93	4540	3573	1899	1102	470	301	107	30	14	0
6	111	4247	3213	1548	849	340	199	64	16	10	0
7	130	4515	2742	1264	617	232	129	40	8	8	0
8	155	4232	2359	1035	450	162	80	26	5	6	0
9	198	4139	1923	786	296	103	45	15	2	5	0
10	918	3911	1224	427	129	44	13	6	1	3	0

This table shows many cells where the proportion of uniques is so high that disclosure control measures can overwhelm the information in the data. It does not make sense to claim to be publishing a statistical summary of the data when in excess of twenty percent of the cases still appear as uniques. This table highlights the extreme dispersion of risk. In the tables the total of all cell entries is 104.6 million. Of that total only 0.73 million were uniques. Hence only about 0.7 percent of personal cell entries fell in cells with uniques. This average clearly hides the variation shown in Table 3.

### **The impact of some rules**

The conventional rules to control disclosure for meshblock data have been of the form that the area must have some minimum population and that the classification must be ‘at the highest level’. To assess rules of this kind we constructed a simple model. Since each cell in a decile table has a minimum n and minimum p, we have considered rules based on each cell in the table. Cells satisfying the size condition have been aggregated to the cell with minimum p, and cells satisfying the p condition have been aggregated to the cell with minimum n. Then risks from Table 3 are applied to the new numbers, and the total number of cases is aggregated across all the permitted cases. Table 4 gives the estimates of the expected number of uniques with the new limits and Table 5 expresses these numbers as a proportion of the number of uniques occurring in the data set.

**Table 4. Expected Number of Uniques With Each Cell Specifying Lower bounds for (n,p)**

		1	2	3	4	5	6	7	8	9	10
		0.002	0.009	0.015	0.026	0.040	0.052	0.066	0.136	0.358	0.950
1	19	734341	724628	650431	561467	435295	363479	262584	151818	77090	13434
2	40	682012	672157	597988	508494	382279	311306	205315	106696	51628	1050
3	60	625253	614995	540488	450607	326743	258566	153659	74956	44105	517
4	77	572032	561487	487970	397792	278190	214751	117420	55824	39038	376
5	93	518925	507641	435286	345141	233152	174491	91308	42529	34787	227
6	111	469800	458405	386166	299394	196548	142019	72323	33487	31067	226
7	130	414127	401035	332045	247850	158327	111448	56792	25801	27697	122
8	155	357796	344402	279080	199160	123944	82561	43674	19888	23212	47
9	198	290414	275688	216270	145063	88164	55578	30981	14445	19159	15
10	918	183835	166800	121558	73161	43868	23837	16995	8792	11874	29

**Table 5. Ratio of Expected Uniques to Initial Levels**

		1	2	3	4	5	6	7	8	9	10
		0.002	0.009	0.015	0.026	0.040	0.052	0.066	0.136	0.358	0.950
1	19	1.000	0.987	0.886	0.765	0.593	0.495	0.358	0.207	0.105	0.018
2	40	0.929	0.915	0.814	0.692	0.521	0.424	0.280	0.145	0.070	0.001
3	60	0.851	0.837	0.736	0.614	0.445	0.352	0.209	0.102	0.060	0.001
4	77	0.779	0.765	0.665	0.542	0.379	0.292	0.160	0.076	0.053	0.001
5	93	0.707	0.691	0.593	0.470	0.317	0.238	0.124	0.058	0.047	0.000
6	111	0.640	0.624	0.526	0.408	0.268	0.193	0.098	0.046	0.042	0.000
7	130	0.564	0.546	0.452	0.338	0.216	0.152	0.077	0.035	0.038	0.000
8	155	0.487	0.469	0.380	0.271	0.169	0.112	0.059	0.027	0.032	0.000
9	198	0.395	0.375	0.295	0.198	0.120	0.076	0.042	0.020	0.026	0.000
10	918	0.250	0.227	0.166	0.100	0.060	0.032	0.023	0.012	0.016	0.000

These tables show how the occurrence of uniques is very resistant to small changes. To halve the number of uniques it is necessary to aggregate all meshblocks with sizes 60 or less with larger meshblocks and do global recodes so that no category proportions are 0.015 or less.

### **Some lessons for disclosure control design**

The objective of this paper is to outline the situations in which SDC problems arise, and to consider the nature of a policy framework to deal with them. The immediate problem is to construct an appropriate framework for the 2006 Census publication outputs. Currently ‘rules’ are used to restrict the nature of output with the objective of meeting the legislative requirements, but apart from the legislated requirements there appear to be no intermediate ‘policy objectives’. In a range of SDC circumstances it can be shown that there are difficulties in meeting the legislated requirements, so a general policy perspective would be valuable. In this section therefore we address some of the issues appropriate to privacy and publication policy.

An attempt to break the SDC tools used in publication involves using personal prior information specific to the intruder, in combination with the published information to create links between the two and extend the inferences the intruder can make about individuals. It is impossible to know the prior information available to the intruder so one cannot in general determine the risk of disclosure from a given publication. This restricts the data provider to establishing some general properties of the data designed to minimise the risk associated with any attempt to use the data to establish respondent characteristics for specific persons. A confidentiality policy should therefore specify so far as possible statistical properties of the data to be published. The properties should include the maximum incidence of disclosure risks prior to application of SDC.

The data agency situation is complicated by the fact that there appear to be no published studies analysing the performance of particular SDC tools applied to tables of counts subject to a variety of methods of attack. It is obvious that data providers will in general want to keep such information in house in any case and the ONS does not even provide detailed information about their SDC procedures to give even further protection. For microdata sets Elliot(2001) and Domingo-Ferrer and Torra(2001) summarise many insights and there has been a great deal of research. For tables Duke-Williams et.al.(1998) give an example of a method of attack which might enable identification of uniques across a set of tables. Dobra and Fienberg (2002) provide other examples. The Duke-Williams method is likely to work better if there are a large number of uniques in the tables. There are a wide range of other methods of attempting to identify uniques in tables.

To simplify our discussion of the issues we will refer to two sets of tables. The source tables are those constructed from the data prior to application of any SDC. The publication tables are the tables after SDC tools have been applied.

Given the uncertainty discussed above, the analysis in this paper would lead us to define at least two properties characterising disclosure risks in the source tables. The first is the probability that an entry in a cell is a unique given by  $E[\mu_{i,j}] = n_i p_{ij} (1 - p_{ij})^{n_i - 1}$  for the  $i$ 'th geographic unit and the  $j$ 'th category. Note we permit the  $p$  values to vary across the geographic units. The second is the probability that a person selected at random from the population appears as a unique in a specific

source table. It is given by  $\frac{\sum_i \sum_j E[\mu_{i,j}]}{\sum_i n_i}$ . For a set of tables it is formed by taking the sum of the

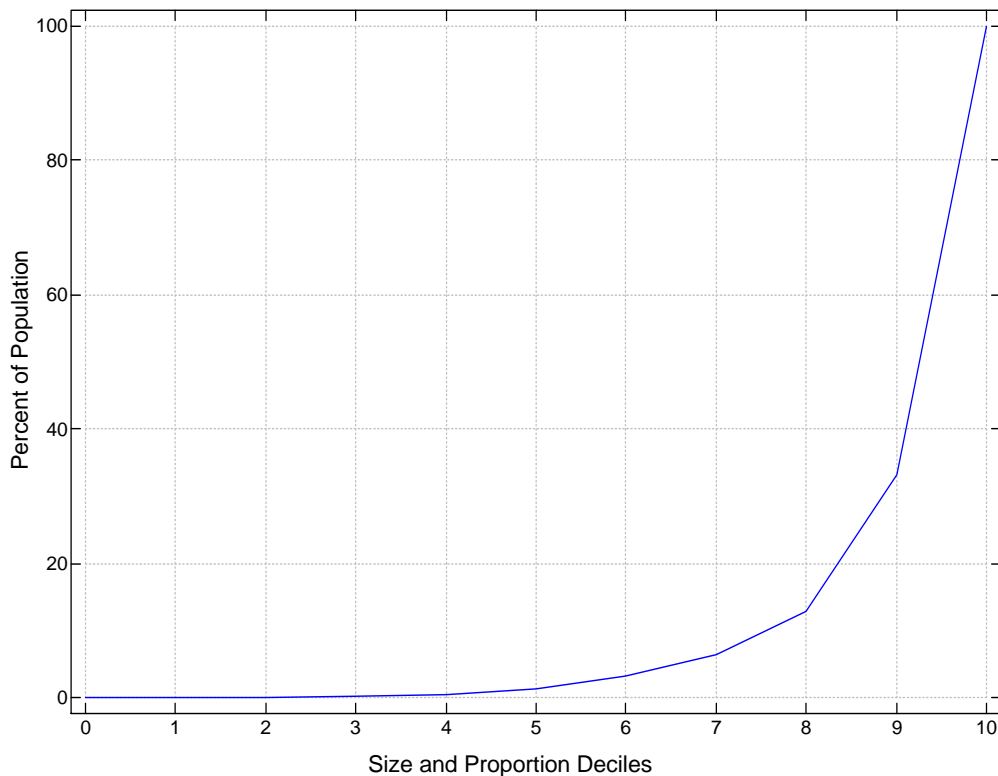
numerators divided by the sum of the denominators for the set of tables. If the probability of any particular person appearing as a unique is sufficiently low, the return from the exercise of attempting to break the SDC tools which yielded the publication tables is likely to be small. These two probabilities are related but they do need to be distinguished. The probability of a unique varies widely across the cells of a table. The first characteristic puts an upper bound on the risk for persons in any subgroup within the table. We have seen that about half of the cells in Table 3 have in excess of ten percent of persons as uniques and some groups have much higher proportions of uniques. A confidentiality policy should seek to ensure that protection is provided to respondents from such small distinctive groups. The second requirement seeks to limit the total number of uniques across all cells. Obviously both the distribution across geographic units and across the categories influences this total. Reducing the number of small units or small categories will reduce the number of cells and the probability of a unique in each cell, so both reduce the total. Steps to reduce the risk on the first property therefore also reduce the risk on the second.

### Uniques for some np combinations – low np cases and per person and per cell risks

A first objective should be to reduce the probability that a person in a category is a unique. The probability is very low for persons in the largest categories and the majority of people are in these groups. Deciles at the top left of Table 1 have the highest Per Cell risks but involve relatively few people. Using Table 2 we can construct a new table by replacing each entry by the total number of persons in a cell together with those not below it or to its right. The entries are then divided by the total number of persons in Table 2. In Figure 13 we show how the diagonal terms in the resulting table increase very slowly. There are less than 1.4 percent of all person entries in the quarter of meshblock cells associated with the quadrant of the decile tables below the median size for both classifications. The cells in this quadrant have larger risks of potential disclosure than most of the remaining cells, and can be controlled with a policy rule based on meshblock size. It will not necessarily reduce the maximal risk since categories with small  $p$  can still generate high risks. Across the 28 variables there are 2.075 million cells in the meshblock tables for this group of cases. The cells have a sum of counts of 1,427,644 and 249,443 uniques, which is more than a third of the total. The diagram shows that a change affecting a relatively small proportion of the population table entries can be associated with large changes in Per Cell risks. Aggregating meshblocks is

therefore a valuable tool to reduce the number of persons with a high risk of appearing as a unique in the tables. In these cells there are many remaining cases which will be twos or threes which should also be considered as sensitive and they would also be eliminated. A systematic policy approach might be to use the model to restrict the incidence of uniques to some stated percentage of cells. Suitable levels might be 10 per cent or less. Because the data suppression methods of SDC also have implications for suppression of marginal entries, lower levels such as 5, 2 or even 1 percent also need to be considered. The empirical analysis in the paper is broadly consistent with the pattern in Figure 3 which achieves 10 percent at  $np = 3.6$  and 5 percent at  $np = 4.5$  and one percent at  $np = 6.5$ . Achieving this would also require some global recoding to raise category probabilities.

A prime lesson of this data is that category classifications should be designed wherever possible to avoid categories with small proportions. This obviously needs to be done sensitively and with careful regard for important subject area distinctions and likely interactions with other variables. At the lowest geographical level there should be significant limits on the minimum category probability, particularly having regard to geographic variability. If it is important for policy reasons to know about the distribution of some category of persons which form a small proportion of the population then it may be appropriate to consider if the census is the appropriate means of collecting and distributing the data.



**Figure 13. Proportion of Population with both  $n$  and  $p$  below upper bound of stated deciles**

Considering the second objective, if a single table is chosen randomly from the set of 28 tables, then the expected number of uniques is 0.65 percent of the population. This puts a different perspective on the problem. It is quite easy to obtain values of the order of 1 percent for this proportion, but to maintain the legislated requirements, a case can be made that it should be at least an order of magnitude lower and probably more. In our study, because there are multiple tables, there would be approximately 100,000 cells containing a unique when this ratio is restricted to 0.1 percent of the population. The high probability categories which generate few uniques mean that the probability that a person chosen randomly from the population is a unique is much lower than

the cell probability. This may be sufficiently low to make the return from a successful intrusion small relative to its cost. For some purposes, this would discourage attempts at intrusion, but for other more malicious attempts the ability to identify even a very few distinctive individuals could be very damaging to the statistical system. Reducing the number of uniques and sensitive cells has a powerful effect by reducing the probability that an individual may be identified as a unique in the population. It would be appropriate to set an upper bound for this proportion since it reduces the maximum return available from a successful intrusion. Since the effect depends on the number of tables being considered within the standard output, one approach would be to set this parameter to restrict the total proportion of the population appearing as uniques across the set of tables. In our example, 0.04 percent would give a total of one percent of the population across the tables of 28 variables.

An approach based on restricting the incidence of uniques for categories within the tables provides a clear framework within which the SDC tools must be effective, and their efficiency can be studied. Until we have much better means of assessing the performance of the SDC tools it will be very difficult to assess whether these limits suggested here are too conservative. In the next two parts of this section we make some comparisons with existing rules used in New Zealand and elsewhere.

### Meshblocks and the values of n

The meshblock system has meant that New Zealand is alone in having a wide variety of sizes of its smallest geographic units. In Australia there are approximately the same number of census collection districts as in New Zealand with a population approaching five times our size. They are much more uniform in size. There are some users of the New Zealand data who would welcome units more uniform in size and large enough to give tighter confidence intervals for observed statistics. Small meshblocks create higher disclosure risks since limited prior information may be adequate to create the restrictions necessary to identify individuals. They also create high proportions of uniques. Because small meshblocks have small numbers of people in them, even a high proportion of uniques does not generate very large numbers of uniques. From the perspective of the total number of uniques eliminating them appears to have a limited effect. However what is important is that these cases are the most likely to be identifiable by an intruder. Leaving persons in small meshblocks increases their risk of identification, relative to those in larger meshblocks. Our analysis has shown that the unit size is an essential parameter in controlling the proportion of uniques, and an overall policy on disclosure control must include measures to influence or control sample size.

In Australia the policy is to have no Collection Districts smaller than 100 though an exception is made for some Aboriginal areas where a bound of 80 is permitted. However the mean size of the Collection Districts is about five times the mean size of meshblocks in New Zealand. In the United Kingdom the smallest units are 100 persons but the mean size is again much larger.

Many users have an investment in the existing meshblock data system and any changes must be designed to accommodate their needs. If the minimum publication unit is larger it should be formed from an aggregate of existing meshblocks. It should be possible to construct an automatic procedure which would construct a set of larger publication units using adjacent meshblocks and their socio-economic characteristics to form new units with a minimum size and a limited range of sizes. We constructed a system of aggregated meshblocks in a manner that left large meshblocks unchanged but aggregated the remainder to minimum sized units. It had a significant impact on the number of uniques and obviously reduced the serious problem of uniques in areas of the table that are sparse. Some of the largest meshblocks should be split. A manual review of the aggregation with local 'experts' would take some time but is not an insuperable task. Statistics New Zealand would obviously need to construct tables for the new publication units for data from previous censuses to simplify the task of users needing to maintain continuity of some analyses. For most users a data concordance linking meshblocks and publication users should be adequate. Where users

insist on wanting to maintain the present units, a simple table lookup system could be devised to subdivide the persons in one of the new publication units into the previous units, but using the publication unit proportions. Such a system might even give a more accurate information base for many analyses. Some statistics like the NZDEP series use existing meshblocks. They would need to be recalculated for the new units. For these statistics the methodology of the index is well established so repeating the calculations with new geographic units should not be an excessively large task.

## Categories and the Values of $p$

The present rules require that tabulations at the meshblock level shall be at the highest level of a classification. This definition is quite general and does not constrain the category probabilities. In other jurisdictions there appear to be similar broad conditions. In the meshblock tables we used the mean cell size was 12.5 persons but there was a huge variation about this and over the quarter of all cells below the median sizes on both classifications the mean cell density was only 0.7 persons per cell. While our examples contain some categories associated with small unspecified groups which do not pose a disclosure risk, most are specific groups with small probabilities. The decision about meshblock size and permitted category probabilities cannot be undertaken separately if the policy is to restrict the percentage of uniques. For example if the policy is to require  $np$  greater than 5 a minimum publication unit of 200 would require  $p$  at least 0.025. A smaller minimum publication unit would require an even larger  $p$ . If the average Australian collection district size was used in New Zealand the minimum  $p$  would be 1 percent thus giving a much wider range of cases than possible with the lower limits we have discussed for New Zealand.

Note that at this stage the suggestion is based on using the category probability and ignoring geographic variability. That does have some risks, but for some sensitive variables there is likely to be a pattern of geographic concentration which will reduce the risks in high density areas and potentially create a problem in the remainder.

There are other approaches to presenting meshblock data which do not involve individual counts. Various descriptive statistics can be used to summarise the data and avoid publication of tables with counts including cells of one or two.

If counts in a set of categories are used, they should be a part of a hierarchic system which enables greater detail where it is appropriate in tables for larger geographic units. Area units average approximately 20 times the current mean meshblock size. Such a unit with about 2000 persons can have a classification with 0.0025 as a minimum category probability and still achieve  $np > 5$ . Where that level of detail is required area units are an appropriate level for the tables.

The approach recommended here has important implications for tables generally. Prior to the 1996 Census tables would only be provided if the mean cell density was at least 4. That is equivalent to a mean  $np$  value for a table of 4. The proposal in this paper is that the rule should apply to the smallest categories since they create the risk of disclosure. This may seem restrictive, but should not be used to completely eliminate tables which would have been published under the old rules. What is required is that the smallest categories be aggregated in a hierarchically consistent manner and the recoded table published. There will undoubtedly be some cases where maintaining consistency of the classification at one of the hierarchic levels is important, but large departures from the target values would need to be avoided.

Providing some control of the category proportions is much more problematic than controlling geographic unit size. The category proportions are subject to geographic heterogeneity, they are dependent on subject matter characteristics and will often have important associations with categories for another variable. However these problems are exactly the reason why some of the more serious disclosure problems arise. Even combining two way tables of variables may provide sufficient information for significant disclosure when the individual belongs to a category which is rare in an area.

The suggestion of using an expected value of  $np$  at the category level is a departure from practice both here and elsewhere. It deliberately provides greater protection for minority groups. In both New Zealand and the UK the most recent censuses have permitted tables with a mean cell size of 1. That seems a particularly inappropriate rule since  $np=1$  is the condition to maximise the number of uniques. Any table for which that is used as the mean for all cells must have many cells with much smaller values of  $np$  and all cells with  $np<0.1$  are subject to the ones in sparse data problem which has been discussed in a previous report.

### Implications for Multi-way Tables

Our analysis shows that it is the category which is the crucial unit for analysis and that should be the level at which targets are set. Controlling  $np$  at the category level automatically modifies the level of detail which may be published for units of widely differing geographical size. How does this transfer to multi-way tables?

The method works because for meshblock tables a good first approximation is independence of the geography and the variable categories. Ultimately however as the use of geographic variability in  $p$  showed, it is estimation of the cell expected value which is important. Multiway tables in which the cell expected values are all 3.5 and either a multinomial or poisson model is appropriate will have about 10 percent of their cells uniques. If the expected value is 1 for all cells the expected number of uniques will be maximised at 36 percent of cells.

Multiway tables with a cell average of 1 will have some cells with very low expected values and some very much larger. The lognormal model used by Skinner and Holmes might be an appropriate model and the Beta-product models of Jackson (2004) are another alternative. In both cases the ones in sparse data problem will arise for some cells, and the risks will clearly be reduced in the mean cell size is larger. The mean size of 4 previously used by Statistics New Zealand might reduce the incidence of uniques for many tables but that would need to be carefully investigated because of the distribution of the expected cell sizes.

### References

Joseph Domingo-Ferrer and Vicenc Torra (2001). Disclosure Control Methods and Information Loss for Microdata. Chapter 5 in Doyle et.al.

Pat Doyle, Julia I Lane, Julems J.M. Theeuwes and Laura M Zayatz (Eds) (2001). Confidentiality, Disclosure and Data Access: Theory and Practical Application for Statistical Agencies. Elsevier Science BV.

Mark Elliot. Disclosure Risk Assessment (2001). Chapter 4 in Doyle et.al. 2001.

Mark Elliot, C.J. Skinner and A. Dale.(1998). Special Uniques, Random uniques and sticky populations: Some counterintuitive effects of geographical detail on disclosure risk. *Research in Official Statistics*. 1(2), pp53-68.

L. Fraser Jackson (2004). An Analysis of Disclosure Risks and Disclosure Protection using Random Rounding with applications to Census tables. Report to Statistics New Zealand.

L. Fraser Jackson, Lisa Corscadden and Irene Zeng (2005). Disclosure Control and Publication Design: When do uniques occur? Paper submitted to ISI 2005.

C.J. Skinner and D.J. Holmes. (1992) Modelling Population Uniqueness. Paper to international seminar on statistical confidentiality. Dublin.

C.E. Weatherburn(1946). A First Course in Mathematical Statistics. Cambridge.

Leon Willenborg and Ton de Waal. Statistical Disclosure Control in Practice.(1996). Springer-Verlag. New York.

Leon Willenborg and Ton de Waal. Elements of Statistical Disclosure Control (2000). Springer-Verlag. New York.

Appendix  
Variables and Category Proportions.

asianind	0.8960	0.0637	0.0403					
urbrural	0.7078	0.0628	0.0839	0.0213	0.1237	0.0005		
depend	0.9494	0.0505	0.0001					
country	0.0006	0.8202	0.0709	0.0052	0.0033	0.0121	0.0239	0.0083
	0.0069	0.0486						
empstat	0.5750	0.0349	0.0574	0.0114	0.3213			
euroind	0.1914	0.7683	0.0403					
famrole	0.2195	0.4570	0.2891	0.0012	0.0004	0.0045	0.0283	
incsrces	0.2673	0.4017	0.2197	0.0557	0.0068	0.0009	0.0001	0.0000
	0.0000	0.0000	0.0000	0.0478				
iwil	0.0375	0.0021	0.0154	0.0108	0.0139	0.0126	0.0091	0.0048
	0.0017	0.0025	0.0092	0.8804				
legmarstat	0.4633	0.3582	0.0286	0.0500	0.0471	0.0528		
maoriind	0.8189	0.1408	0.0403					
occ	0.2847	0.0642	0.0510	0.0579	0.0649	0.0368	0.0389	0.0386
	0.3630							
offlang	0.0018	0.7676	0.0352	0.0002	0.1080	0.0057	0.0150	0.0201
	0.0464							
pacind	0.8977	0.0620	0.0403					
qualfield	0.6311	0.0135	0.0065	0.0399	0.0133	0.0087	0.0308	0.0239
	0.0403	0.0328	0.0089	0.0097	0.0120	0.0001	0.0000	0.0055
	0.0089	0.1141						
quallevel	0.6311	0.0284	0.0369	0.0147	0.0559	0.0535	0.0247	0.0262
	0.0001	0.0001	0.0054	0.0089	0.1141			
qual	0.4105	0.1041	0.0759	0.0420	0.0006	0.0438	0.0284	0.0369
	0.0147	0.0559	0.0535	0.0247	0.1090			
religion1	0.2751	0.0108	0.5390	0.0100	0.0058	0.0014	0.0146	0.0036
	0.0039	0.1358						
seekwork	0.2837	0.2434	0.4729					
sex	0.4878	0.5122						
tenure	0.6219	0.3252	0.0046	0.0483				
totinc	0.2314	0.0323	0.0644	0.0903	0.1057	0.0652	0.0558	0.0575
	0.0834	0.0490	0.0438	0.0183	0.0167	0.0862		
trawrk	0.2735	0.0563	0.2239	0.0442	0.0209	0.0147	0.0041	0.0048
	0.0113	0.0256	0.0037	0.3170				
unpaid	0.8311	0.1689						

ind	0.2268	0.0382	0.0009	0.0599	0.0016	0.0278	0.0266	0.0558
	0.0215	0.0178	0.0062	0.0139	0.0521	0.0159	0.0339	0.0376
	0.0111	0.0172	0.3352					
wkfstat	0.9386	0.0124	0.0261	0.0229				
socmar	0.2364	0.3498	0.0805	0.0182	0.1732	0.0210	0.0315	0.0398
	0.0496							
age5r	0.0725	0.0765	0.0778	0.0710	0.0642	0.0660	0.0748	0.0796
	0.0764	0.0674	0.0632	0.0487	0.0414	0.0342	0.0316	0.0253
	0.0164	0.0130						